

**Tilburg University**

## **Bounds on Quantiles in the Presence of Full and Partial Item Nonresponse**

Vazquez-Alvarez, R.; Melenberg, B.; van Soest, A.H.O.

*Publication date:*  
1999

[Link to publication in Tilburg University Research Portal](#)

*Citation for published version (APA):*

Vazquez-Alvarez, R., Melenberg, B., & van Soest, A. H. O. (1999). *Bounds on Quantiles in the Presence of Full and Partial Item Nonresponse*. (CentER Discussion Paper; Vol. 1999-38). Econometrics.

### **General rights**

Copyright and moral rights for the publications made accessible in the public portal are retained by the authors and/or other copyright owners and it is a condition of accessing publications that users recognise and abide by the legal requirements associated with these rights.

- Users may download and print one copy of any publication from the public portal for the purpose of private study or research.
- You may not further distribute the material or use it for any profit-making activity or commercial gain
- You may freely distribute the URL identifying the publication in the public portal

### **Take down policy**

If you believe that this document breaches copyright please contact us providing details, and we will remove access to the work immediately and investigate your claim.

# Bounds on Quantiles in the Presence of Full and Partial Item Nonresponse<sup>1</sup>

Rosalia Vazquez Alvarez

Bertrand Melenberg

Arthur van Soest

Corresponding author:

Rosalia Vazquez Alvarez

Department of Econometrics

Tilburg University

P.O.Box 90153

5000 LE Tilburg

The Netherlands

E-mail: r.vazquez-alvarez@kub.nl

## Abstract

Microeconomic surveys are usually subject to the problem of item nonresponse, typically associated with variables like income and wealth, where confidentiality and/or lack of accurate information can affect the response behavior of the individual. Follow up categorical questions can reduce item nonresponse and provide additional partial information on the missing value, hence improving the quality of the data. In this paper we allow item nonresponse to be non-random and extend Manski's approach of estimating bounds to identify an upper and lower limit for the parameter of interest (the distribution function or its quantiles). Our extension consists of deriving bounding intervals taking into account all three types of response behavior: full response, partial (categorical) response and full nonresponse. We illustrate the theory by estimating bounds for the quantiles of the distribution of amounts held in savings accounts. We consider worst case bounds which cannot be improved upon without additional assumptions, as well as bounds that follow from different assumptions of monotonicity.

**Key words:** item nonresponse, bracket response, bounds and identification.

**JEL Classification:** C13, D31

---

<sup>1</sup> We are grateful to Rob Alessie and Bas Donkers for their helpful comments. Gauss programs used in this paper are available from the corresponding author.

# 1 Introduction

The aim of economic surveys is to collect data to provide the possibility to study social and economic trends in the population of interest. For example, at a microeconomic level, important questions in household surveys focus on the savings behavior of the household, the distribution of wealth, and the distribution of income. Longitudinal studies such as the Panel Study of Income Dynamics (PSID), the Health and Retirement Survey (HRS), and the Asset and Health Dynamics Among the Oldest Old (AHEAD) are usually thought of as high quality data providers for microeconomic studies. These panels, however, are subject to the problem of missing data. Non-negligible missing data occurs when a significant amount of individuals in the panel gives no answers to any of the questions in the survey (unit nonresponse) or provides answers to some of the questions in the survey, but not all (item nonresponse). Item nonresponse is usually encountered in questions where individuals are asked to disclose their income, earnings or wealth with an exact amount. Nonrandom item nonresponse complicates the use of the data since it will generally result in a sample which is not representative of the population of interest. If not accounted for appropriately, nonrandom nonresponse can bias the results of studies which, for example, try to explain why people save, analyze the income distribution, or forecast accumulation of household wealth.

Item nonresponse can be treated in two stages: at the data collection level and at the estimation stage. At the data collection level, the problem of item nonresponse might be reduced by adding follow up questions in which initial non-respondents only need to reveal some categorical information about their savings, wealth or incomes. This technique to reduce the problem of nonresponse, is motivated by the claim that certain cognitive factors, such as the belief that the interviewer requires very precise information and/or confidentiality reasons, may explain why people are more reluctant to disclose information on assets and incomes, compared to other social and economic variables (see, for example, Hurd et. al, 1997). Juster et al. (1997) examined the 1993 wave of the HRS panel and showed how categorical questions can dramatically cut nonresponse rates in questions related to assets; for example, answers on savings accounts showed an initial nonresponse rate of 28% but a combination of categorical questions reduced this to only 8% of full nonresponse. Juster et al. (1997) also emphasize that categorical questions may have secondary effects on the response behavior: for example, one might think that individuals who answer in brackets at early stages of the interview will have a tendency to answer in brackets at later stages, resulting in loss of exact information. Opposite to this, however, they observe that individuals learn that a good approximation to asset values is sufficient, so that at later stages in the interview, they immediately provide rounded amounts, thus avoiding the lengthier categorical questions. Hurd et al. (1997) discuss the role of

categorical questions in the panel data AHEAD, and the effectiveness of this type of questions on reducing item nonresponse. They also point out how the design of the categorical questions can influence response and can lead to biased estimates of location measures of interest (the so called anchoring effect).

Initial non-respondents who either answer ‘don’t know’ or ‘refuse’, when asked about a specific amount, often face one of two possible types of categorical questions. The first is range cards, where the individual is shown a complete range of categories partitioning all possible amounts, so that the respondent can choose the interval containing the amount in question; range cards also include the possibility to answer ‘don’t know’ or ‘refuse’. The second type of categorical questions is known as unfolding brackets; here, initial non-respondents are asked to answer ‘yes’, ‘no’, ‘don’t know’ or ‘refuse’ to a question such as ‘*is the amount \$x or more?*’; this question is asked various times - usually three to four times with different values of \$x. If the respondent ends by answering with an inconclusive statement, such as ‘don’t know’ or ‘refuse’, no further follow up questions are asked. Both methods have their advantages and disadvantages. The main advantage of unfolding brackets is that, although an initial non-respondent can end up answering with ‘don’t know’ or ‘refuse’ at some point of the unfolding bracket design, it is likely that, before this happens, he or she will reveal partial information about the amount in question; on the other hand, initial non-respondents that face a range card may often choose the option ‘don’t know’, without providing any partial information at all. If participants answer questions over the telephone, unfolding brackets is the only possible design in terms of categorical questioning, since range cards cannot be used in telephone interviews. A problem with unfolding brackets is that they are more time consuming than range cards. Moreover, range cards typically allow for more choices of categories than unfolding brackets. Finally, unfolding brackets lead to ‘the anchoring effect’, meaning that the order in which category bounds are asked affects the answer of the respondent (see Hurd et al., 1997); this order plays no role in range cards where all the bounds are given simultaneously.

Once the raw data is collected, with or without the use of categorical questions, item nonresponse remains a potential problem. Ideally, inference requires a full set of data representative of the population under study; in case of item nonresponse, some people provide full information, some may provide partial information and some no information at all. One way to deal with this is simply to ignore nonresponse units, and conduct inference using only those individuals that provide full information. Completely random item nonresponse or exogenous selection is the underlying assumption that makes this practice feasible. An alternative is to impute the missing values; imputation allows the researcher to obtain a full set of data while using all the available information in the sample. A conventional way to impute missing observations is to use a hot-deck approach. This methodology assumes that the complete sample

can be used as a pool of donors of information on the missing value for nonresponse individuals. The underlying assumption is that, although item nonresponse can be nonrandom, item nonrespondents are not different from respondents that have similar characteristics, given by a set of variables  $X$ . The observed values on the variables of interest from respondents can be used to impute those of non-respondents with similar values of  $X$ . In other words, item nonresponse is assumed to be random conditional on  $X$ . Juster et al. (1997) show that hot-deck imputation can also benefit from follow up categorical questions. Comparing conventional hot-deck imputation with hot-deck imputation using bracket response they found that conventional hot-deck imputation understates population estimates of non-housing wealth by at least 19%.

Since the seminal work by Heckman (see Heckman, 1979, for example), it is well known that if item nonresponse is non-random, simply deleting item non-respondents can lead to a selection bias. To solve this selection problem one could use a selectivity model that takes account of selectivity bias and avoids the assumption of complete (or conditional) random item nonresponse. This leads to a class of parametric and semiparametric models which generalize the original Heckman model. These models typically impose some parametric and semiparametric restrictions on the conditional distribution of the variable of interest  $Y$  given covariates  $X$ , and on the item nonresponse mechanism.

Although selectivity models and imputation procedures are well established methods to deal with item nonresponse, both procedures share the problem that they require additional assumptions. Since the early 1990's Charles Manski has put forward a new approach to deal with censored data in the form of item nonresponse which avoids such assumptions; see Manski (1989, 1990, 1994, 1995), but also Heckman (1990). This approach is usually concerned with the full conditional distribution function of a variable  $Y$  given a specified value of some vector of variables  $X$ . The idea is that, without additional assumptions, the parameter of interest is identified up to a bounding interval. Item nonresponse is allowed to be nonrandom. Let  $\delta$  be a binary random variable that takes the value one if  $y$  is observed and the value zero otherwise, so that each member of the population is characterized by  $(Y, \delta, X)$ . A random sample from the population will reveal  $(\delta, x)$  for all observations, while  $y$  will only be observed if  $\delta=1$ . It is not possible to identify the distribution function of  $(Y, \delta, X)$ , since the censored-sampling process is uninformative with respect to the distribution function of the sub-population with  $\delta=0$ . Prior information can be used to derive bounding intervals around the distribution function of the whole population. For example, the fact that the range of the distribution function for non-respondents is in the  $[0,1]$  interval, leads to the so called worst case bounds, where the distance between the upper and the lower bound is driven by the conditional probability of nonresponse (see Manski, 1995); these bounds cannot be improved upon without making additional assumptions. Nonparametric assumptions such as monotonicity or exclusion restrictions, can

lead to narrower sets of bounding intervals on the unknown parameter of interest that improve on the worst case bounds (see Vazquez et al. (1999) for an application to earnings).

In this paper we extend Manski's approach by allowing the possibility of initial non-respondents being directed to a categorical question where they can reveal partial information. Thus three sub-populations are distinguished: full respondents, bracket respondents and full non-respondents. We show how worst case bounds on the unknown distribution function and quantiles can be narrowed by taking into account bracket response. We apply this to estimate bounds on the quantiles of the distribution of savings in a representative sample of the Dutch population. In this sample, the initial item nonresponse rate is approximately 40%, but since initial non-respondents are routed to a range card categorical question, the final item nonresponse rate drops to only about 12% of the total sample.

The remainder of this paper is organized as follows. Section 2 reviews Manski's (1995) worst case bounds and derives a new set of worst case bounds taking into account categorical information. In this section, we also explore the assumption of monotonicity when the worst case set of bounds depends on three levels of response. Section 3 describes the data used in the empirical illustration. Section 4 explains the estimation technique and discusses the empirical results. Section 5 concludes.

## 2 Theoretical framework

### 2.1 Worst Case bounds on the distribution function

In this section we start by reviewing Manski's (1995) worst case set of bounds for the conditional distribution function of a variable  $Y$ , at a given  $y \in \mathbb{R}$ , and given  $X = x \in \mathbb{R}^p$ . We assume that there is no unit nonresponse in the sample, no item nonresponse for the variables in  $x$ , and no measurement error such as under or over reporting the value of  $Y$ . Let the dummy variable  $\delta$  model item nonresponse, i.e  $\delta=1$ , if  $Y$  is observed and zero otherwise. With this,  $F_{Y|x}(y)$ , the conditional distribution function for the whole population, can be expressed as follows:

$$F_{Y|x}(y) = F_{Y(x,\delta=1)}(y)P(\delta=1|x) + F_{Y(x,\delta=0)}(y)P(\delta=0|x) \quad (1)$$

Under the assumptions that we have made,  $F_{Y(x,\delta=1)}(y)$  is identified for all  $x$  in the support of  $X$ .  $F_{Y(x,\delta=1)}(y)$  can be estimated using some nonparametric estimator. Similarly,  $P(\delta=1|x)$  and  $P(\delta=0|x)$  are identified and can be consistently estimated, since we are assuming

complete response on  $\delta$  and  $X$ . If we assume that  $\delta$  is independent of  $Y$  conditional on  $X$ , then  $F_{Y(x,\delta=1)}(y) = F_{Y(x,\delta=0)}(y)$  and all expressions in the right hand side of (1) are identified; this is the assumption of exogenous selection. In general, however,  $\delta$  can be related to  $Y$ , and  $F_{Y(x,\delta=0)}(y)$  is then not identified, so that  $F_{Y|x}(y)$  is not identified either. With no additional assumptions, all we know about the distribution function for the nonresponse sub-population is that  $0 \leq F_{Y(x,\delta=0)}(y) \leq 1$ . Applying this to (1) gives

$$F_{Y(x,\delta=1)}(y)P(\delta=1|x) \leq F_{Y|x}(y) \leq F_{Y(x,\delta=1)}(y)P(\delta=1|x) + P(\delta=0|x) \quad (2)$$

This expression shows Manski's (1995) basic worst case upper and lower bounds. The width between these bounds is equal to  $P(\delta=0|x)$ . The larger the probability of nonresponse, the wider the interval between upper and lower bound and the less information we obtain about the unknown distribution function. Unless one makes additional assumptions or has additional information on the item non-respondents, these bounds cannot be improved upon.

If the survey allows initial non-respondents to disclose partial information on the dependent variable with a categorical question - which we assume to be of a range card type -, the sample can be split into three sub-populations. Using the categorical information leads to a new set of bounds that can be more informative about the unknown distribution function than expression (2); we call these new bounds also the worst case set of bounds, because, similar to expression (2), the new set does not require additional assumptions.

Allowing initial non-respondents to disclose partial information implies that response can be at three levels according to two observed dummy variables  $\delta_1$  and  $\delta_2$ :

$$\begin{aligned} \delta_1 &= 1 \quad \text{if full response on } Y \\ \delta_2 &= 1 \text{ and } \delta_1 = 0 \quad \text{if response in bracket} \\ \delta_2 &= 0 \text{ and } \delta_1 = 0 \quad \text{if full nonresponse} \end{aligned} \quad (3)$$

The partition in (3), leads to the following expression replacing (1):<sup>2</sup>

---

<sup>2</sup> For simplicity, and according to the range card categorical questions used in our empirical example, from this point onwards we assume the brackets are the same for all sample observations; this would be easy to generalize, however.

$$F(y|x) = F(y|\delta_1=1,x)P(\delta_1=1|x) + F(y|\delta_1=0,\delta_2=1,x)P(\delta_1=0,\delta_2=1|x) + F(y|\delta_1=0,\delta_2=0,x)P(\delta_1=0,\delta_2=0|x). \quad (4)$$

The censored-sampling process does not identify all elements in (4) as it is not informative about  $F(y|\delta_1=0, \delta_2=0,x)$ , and not fully informative about  $F(y|\delta_1=0, \delta_2=1,x)$ ; all we know about these two expressions is that

$$0 \leq F(y|\delta_1=0,\delta_2=0,x) \leq 1$$

*and*

$$F(L(y)|\delta_1=0,\delta_2=1,x) \leq F(y|\delta_1=0,\delta_2=1,x) \leq F(U(y)|\delta_1=0,\delta_2=1,x) \quad (5)$$

where  $L(y)$  and  $U(y)$  are the bounds of the brackets containing  $y$ , i.e.  $L(y) \leq y \leq U(y)$ . For example, if we have bounds  $f.0,00 - f.25,000$ ,  $f.25,000 - f.50,000$ , and  $\geq f. 50,000$ , then both  $F(50,000|\delta_1=0,\delta_2=1,x)$  and  $F(25,000|\delta_1=0,\delta_2=1,x)$  are identified by the data, because partial respondents indicate whether their  $y$  value is between the specific values of  $f. 25,000$  and  $f.50,000$ . For values of  $y$  above or equal to  $f.50,000$ , for example, we know that  $F(y|\delta_1=0,\delta_2=1,x) \geq F(50,000|\delta_1=0,\delta_2=1,x)$ . Applying (5) to (4) leads to the following worst case set of bounds:

$$F(y|\delta_1=1,x)P(\delta_1=1|x) + F(L(y)|\delta_1=0,\delta_2=1,x)P(\delta_1=0,\delta_2=1|x) \leq F(y|x) \leq F(y|\delta_1=1,x)P(\delta_1=1|x) + F(U(y)|\delta_1=0,\delta_2=1,x)P(\delta_1=0,\delta_2=1|x) + P(\delta_1=0,\delta_2=0|x) \quad (6)$$

In (6) the width between upper and lower bounds is equal to

$$P(\delta_1=0|x) + P(\delta_1=0,\delta_2=1|x) [F(U(y)|\delta_1=0,\delta_2=1,x) - F(L(y)|\delta_1=0,\delta_2=1,x) - 1], \quad (7)$$

where  $P(\delta_1=0|x) = P(\delta_1=0,\delta_2=1|x) + P(\delta_1=0,\delta_2=0|x)$  is equal to the initial nonresponse probability not considering categorical questions. Clearly, (7) is almost equal to  $P(\delta=0|x)$ , but if a nonzero percentage of the population answers to the bracket question, and if the brackets are not too large, the expression  $[F(U(y)|\delta_1=0,\delta_2=1) - F(L(y)|\delta_1=0,\delta_2=1) - 1]$  will be negative and the bounds in (6) will be sharper than those in (2). Therefore, using follow up categorical questions



generally will improve upon Manski's original worst case bounds.

## 2.2 Bounds on the distribution function and monotonicity

Manski (1995) employs the concept of monotonicity to illustrate the consequences of imposing weak additional assumptions when dealing with censored data. Vazquez et al. (1999) illustrate this empirically and show that the use of a monotonicity assumption leads to narrower bounds than the worst case bounds. In these studies, bracket response was not an issue and with only two populations (item respondents and item non respondents), the concept of monotonicity implied only three possible relations between the two sub-population distributions:  $F(y/\delta=0, x) = F(y/\delta=1, x)$ ,  $F(y/\delta=0, x) \leq F(y/\delta=1, x)$  and  $F(y/\delta=0, x) \geq F(y/\delta=1, x)$ . The choice among these three can be made on the basis of prior beliefs on response behavior.

In the presence of bracket response, the three sub-populations lead to 28 possible relations among the three distribution functions (see Appendix A). Each of these implies a different monotonicity assumption. Not all 28 relations are equally plausible, and many of them will appear to be inconsistent with the full response and bracket response data in our empirical example. We will, therefore, only derive the bounds under three monotonicity assumptions which seem to be a plausible interpretation of our data<sup>3</sup>; the other 25 cases can be derived in a similar way.

We will use the following short hand notation:

*Original notation:*

$$F(y/\delta_1=1, x)$$

$$F(y/\delta_1=0, \delta_2=1, x)$$

$$F(y/\delta_1=0, \delta_2=0, x)$$

$$F(U(y)/\delta_1=0, \delta_2=1, x)$$

$$F(L(y)/\delta_1=0, \delta_2=1, x)$$

$$P(\delta_1=1/x)$$

$$P(\delta_1=0, \delta_2=1/x)$$

$$P(\delta_1=0, \delta_2=0/x)$$

$$P(01)+P(00)=P(\delta_1=0/x)$$

*Shorthand notation:*

$$F_1$$

$$F_{01}$$

$$F_{00}$$

$$F(u)$$

$$F(l)$$

$$P(1)$$

$$P(01)$$

$$P(00)$$

$$P(0)$$

---

<sup>3</sup> See Pages 10 to 13 for a discussion.

In terms of this notation, our three choices from Appendix A are  $F_{00} \leq F_I \leq F_{0I}$ ,  $F_I \leq F_{0I} \leq F_{00}$ , and  $F_I \leq F_{0I}$ .

The inequality  $F_I \leq F_{0I}$  can partially be checked from the data since the data identifies  $F_I$  at all values of  $y$  and  $F_{0I}$  at all bracket bounds. In Section 4 we will show that in the empirical example it is reasonable to impose  $F_I \leq F_{0I}$  (and not  $F_I = F_{0I}$  or  $F_I \geq F_{0I}$ ). This assumption implies that at each value of  $y$  (savings, say, as in our empirical example), the conditional probability of savings for full respondents is below that of bracket respondents. Thus, full respondents, on average, are higher savers than bracket respondents. A reason for this monotonicity assumption could be that higher savers keep better records of their savings and can track the exact amount more easily. In this view bracket respondents do not know the exact amount, but once they are routed to a question where exact knowledge is not important, they have no problem on disclosing partial information.

The relation  $F_I \leq F_{0I} \leq F_{00}$  implies that, in addition to  $F_I \leq F_{0I}$ , full non-respondents are those who tend to have the lowest savings. This could be explained from a similar lack of information argument. If respondents are better informed the higher their savings are, people with low savings will more often not even know enough to determine in which bracket their savings are. On the other hand, this inequality is in contrast with the often given argument that full non-respondents tend to have high savings and refuse to reveal the amount due to privacy concerns.<sup>4</sup>

The final monotonicity assumption we consider is  $F_{00} \leq F_I \leq F_{0I}$ . This implies that the highest savers in the population tend to be full non-respondents. This assumption implies that initial non-respondents consists of two groups. On the one hand for low savers confidentiality is not an issue but lack of exact information prevents them from answering the initial question. They have no problem providing partial information in brackets. On the other hand, there is a group of initial non-respondents with high savings who refuse to provide any information on their savings amount for confidentiality reasons.

Notice that the monotonicity assumption given by  $F_I \leq F_{0I}$  is implied by the other two types of monotonicity. Therefore, we refer to  $F_I \leq F_{0I}$  as the weak monotonicity assumption, since it assumes nothing about the distribution function of full non-respondents. On the other hand,  $F_{00} \leq F_I \leq F_{0I}$  and  $F_I \leq F_{0I} \leq F_{00}$  are non-nested. We will refer to them as Monotonicity 1 and Monotonicity 2, respectively.

---

<sup>4</sup>This suggests that it may be worthwhile to distinguish between initial non-respondents who do not know the answer and those who refuse to give the answer. We do not pursue this here, since the distinction is not present in the data of our empirical example.

## The Weak Monotonicity assumption

The weak monotonicity assumption

$$F_1 \leq F_{01} \quad (8)$$

implies

$$0 \leq F_{00} \leq 1$$

$$\text{and} \quad (9)$$

$$\max[F_1, F(l)] \leq F_{01} \leq F(u)$$

Applying (9) to (4) leads to the following set of upper and lower bounds

$$\begin{aligned} F_1 P(1) + \max[F_1, F(l)] P(01) \\ \leq F_{y|x} \\ F_1 P(1) + P(00) + F(u) P(01) \end{aligned} \quad (10)$$

The width between upper and lower bounds in (10) equals

$$P(0) + P(01)[F(u) - \max[F_1, F(l)] - 1] - P(00) \quad (11)$$

Comparing (11) to (7) shows that the bounds in (10) will be sharper than bounds in (6) if  $F_1 > F(l)$ . If the monotonicity condition is satisfied, i.e.  $F_1 \leq F_{01}$ , we can still have  $F_1 > F(l)$  for values of  $y$  not too close to the lower limit of their bracket, showing that weak monotonicity can be useful.

## Monotonicity 1

The monotonicity assumption

$$F_1 \leq F_{01} \leq F_{00} \quad (12)$$

implies

$$\begin{aligned} \max[F_1, F(l)] &\leq F_{00} \leq 1 \\ &\text{and} \\ \max[F_1, F(l)] &\leq F_{01} \leq F(u) \end{aligned} \tag{13}$$

and applying (13) to (4) leads to the bounds

$$\begin{aligned} F_1 P(1) + \max[F_1, F(l)] P(0) \\ \leq F_{y|x} \leq \\ F_1 P(1) + F(u) P(01) + P(00) \end{aligned} \tag{14}$$

The width between upper and lower bounds in (14) equals

$$P(0) + P(01)[F(u) - \max[F_1, F(l)] - 1] - P(00)\max[F_1, F(l)] \tag{15}$$

The lower bound in expression (14) differs with respect to the lower worst case bound in (6) by  $\max[F_1, F(l)][P(00) + P(01)] - F(l)P(01)$ . If  $F_1 > F(l)$  the difference between the lower bounds equals  $P(00)F_1 + P(01)[F_1 - F(l)]$ ; otherwise the difference is  $F(l)P(00)$ . In both cases it is positive, so that bounds in (14) are sharper than those in (6). We can also compare bounds under the Weak Monotonicity assumption with bounds based on Monotonicity 1; again, their lower bounds differ by  $\max[F_1, F(l)]P(00)$  so that bounds in (14) are sharper than those in (10) as long as  $P(00)$  is positive.

## Monotonicity 2

The monotonicity assumption

$$F_{00} \leq F_1 \leq F_{01} \tag{16}$$

implies

$$0 \leq F_{00} \leq F_1$$

and

$$\max[F_1, F(l)] \leq F_{01} \leq F(u)$$

(17)

Applying (17) to (4) leads to the bounds

$$F_1 P(1) + \max[F_1, F(l)] P(01)$$

$$\leq F_{y|x} \leq$$

(18)

$$F_1 [P(1) + P(00)] + F(u) P(01)$$

The width between upper and lower bounds in (18) is

$$P(0) + P(01) [F(u) - \max[F_1, F(l)] - 1] - P(00) [1 - F_1]$$

(19)

The lower bound differs from that in (6) only if  $F_1 > F(l)$ , whereas the difference with the upper bound in (6) equals  $P(00)[1 - F_1]$ . The bounds in (18) are thus sharper than those in (6) as long as  $F_1 > 0$ ; this can be seen by comparing expression (7) and (19). Expression (10) - bounds under Weak Monotonicity - and expression (18) have the same lower bound, and since the upper bound of expression (10) is identical to the upper bound in (6), the total difference between (10) and (18) equals  $P(00)[1 - F_1]$ . Whether bounds in (18) are narrower than bounds in (14) cannot be determined from a theoretical point of view, since Monotonicity 1 and Monotonicity 2 are non-nested.

## 2.3 Bounds on Quantiles

Distributions for variables like income, savings, etc., are often described in terms of quantiles. For  $\alpha \in [0, 1]$ , the  $\alpha$ -quantile of the conditional distribution of  $Y$  given  $X=x$ , is the smallest number  $q(\alpha, x)$  that satisfies  $F_Y[q(\alpha, x)] \geq \alpha$ :

$$q(\alpha, x) \equiv \inf \{y: F_{Y|x}(y) \geq \alpha\} \quad (20)$$

For  $\alpha > 1$ , we set  $q(\alpha, x) = \infty$ , and for  $\alpha < 0$ ,  $q(\alpha, x) = -\infty$ . The bounds for the quantiles follow from those for the distribution functions by ‘inverting’ (2), (6), (10), (14) and (18): these can be written as

$$lb(y, x) \leq F_{Y|x}(y) \leq ub(y, x) \quad (21)$$

for appropriate choices of  $lb(y, x)$  and  $ub(y, x)$ , all of them non-decreasing functions of  $y$ . Inverting this gives:

$$\inf \{y: lb(y, x) \geq \alpha\} \geq \inf \{y: F_{Y|x}(y) \geq \alpha\} \geq \inf \{y: ub(y, x) \geq \alpha\} \quad (22)$$

Plugging in the bounds on the distribution function in (2), (6), (10), (14) and (18) in (22) thus yields bounds on the conditional quantiles of  $Y$ . Each set of upper and lower bounds can be represented in terms of the distribution function, in which case the percentage of nonresponse is interpreted as the vertical width between bounds in a graph of the distribution function, or by means of the quantiles, in which case nonresponse is reflected by the horizontal width between bounds in the same graph.

### 3 The Data

We use the 1993 wave of the CentER Panel. This panel is a joint venture between the VSB foundation and CentER for Economic Research (Tilburg University) and aims at providing a better understanding of household savings and household financial decision making in The Netherlands (see Nyhus (1996) for more detailed information). We will illustrate the usefulness of the bounds derived above with an empirical example concerning savings of Dutch individuals.

The panel, dating from 1992, collects economic, sociological and psychological information from approximately 3000 households in the Netherlands; the participants are members of the surveyed households of age 16 or more. The panel is made up of two different sub-panels, the Representative sub-panel and the High Income sub-panel. The Representative sub-panel contains approximately 2000 households and is designed to be representative of the Dutch population. The High Income sub-panel, with approximately 1000 households, should

represent units in the top decile of the income distribution. In both sub-panels data are collected by means of a computerized system. We restrict attention to the Representative sub-panel only, so that our initial sample contains 2794 individuals.

The survey contains five different sections. One of these sections, named ‘assets and loans’, provides information about individuals’ assets such as the value of their shares, housing wealth and savings accounts. We will consider the variable savings. As many other panels, the CentER Panel shows a significant percentage of nonresponse for this variable. Questions on savings are designed such that initial non-respondents are routed to a range card type of categorical question. Initially, participants are asked how many savings accounts they possess. All 2794 individuals from the Representative sub-panel answered this question; 2039 individuals report to have zero savings account and the remaining 755 have one or more of such accounts.<sup>5</sup> Our empirical example concerns the amount of savings in the first savings accounts of these 755 individuals.<sup>6</sup>

**Table 1: Means (standard deviations) and Percentages (standard errors) for a selection of social and economic variables.**

	<b>Representative sub-sample</b>	<b>Units with zero savings accounts</b>	<b>Units with at least one savings account</b>
<b>Units</b>	2794	2039	755
<b>Age</b>	44.5 (16.3)	44.2 (16.2)	45.2 (16.7)
<b>% Male</b>	51 (0.9)	46 (1.1)	64 (1.7)
<b>Family size</b>	2.62 (1.31)	2.70 (1.34)	2.40 (1.21)
<b>Education level*</b>	2.30 (0.75)	2.24 (0.73)	2.46 (0.81)
<b>% of house owners</b>	60 (0.9)	58 (1.0)	66 (1.7)
<b>% with savings accounts</b>	27 (0.8)	0.00	100
<b>Number of savings accounts</b>	0.43 (1.60)	0.00	1.42 (1.70)

\*The variable education level measures the educational achievement of individuals on a scale from 1 to 4, where 1 indicates minimum schooling (primary education) and 4 indicates an advanced university degree.

<sup>5</sup> The survey distinguishes between savings accounts linked to a checking account with the postal bank and other savings and deposit accounts. We only consider the latter.

<sup>6</sup> The 755 individuals represent a total of 686 households. Thus 9.1% of our sample belong to the same household as other individuals in the sample. Given this small percentage we continue our analysis assuming independence between savings and response behavior of all individuals in the sample.

Table 1 shows summary statistics for some socio-economic variables for the participants in the Representative sub-panel. On average, those who hold one savings account are older than those who hold zero savings account and belong to smaller households. Holders of savings accounts have higher educational achievement than non-holders and are more likely to own a house. Females are less likely to hold a savings account than males.

The initial question on the first savings account (asked only to individuals with at least one savings account) is as follows,

*‘...What was the balance of your 1<sup>st</sup> account on 31 December 1992?*

*1 - ‘any amount’ in Dutch guilders*

*Don’t know...’*

A total of 455 individuals answered this question with a specific amount. The minimum amount reported was *f*.1 and the maximum *f*.228,767. The median for this group was *f*.6,000 with a standard deviation around the mean equal to *f*.29,494. 300 individuals answered ‘Don’t know’, implying a 39.7% initial nonresponse rate. This latter group was routed to the following range card type categorical question.

*‘...Into which of the categories mentioned below did the balance of your 1<sup>st</sup> savings account go on 31 December 1992?’*

Each initial non-respondent could choose one of the intervals mentioned in Table 2 or the ‘Don’t know’ option. Out of 300 initial non-respondents, 207 gave an answer in one of the intervals. The remaining 93 are full non-respondents. Thus the range card question reduces full nonresponse from 39.7% to 12.3% so that it seems worthwhile to take the range card information into account.

Table 2 shows the distribution of bracket respondents; 38% of the initial 300 non-respondents report that their savings are in one of the lowest two categories: this corresponds to 55% of the 207 bracket respondents. Since the median for full respondents is *f*.6,000, this already suggests that, relative to full respondents, the bracket response individuals might tend to have lower savings.



**Table 2: Distribution of range card answers of initial non-respondents.**

<b>Category</b>	<b>Limits for each category</b>	<b>Percentage of respondents</b>
Category 1	less than fl. 2,000	22 %
Category 2	f. 2,000 - f. 5,000	16 %
Category 3	f. 5,000 - f. 10,000	11 %
Category 4	f. 10,000 - f. 15,000	6.7 %
Category 5	f. 15,000 - f. 20,000	3.3 %
Category 6	f. 20,000 - f. 25,000	3.0 %
Category 7	f. 25,000 - f. 30,000	1.3 %
Category 8	f. 30,000 - f. 40,000	2.0 %
Category 9	f. 40,000 - f. 50,000	0.3 %
Category 10	f. 50,000 - f. 100,000	1.3 %
Category 11	f. 100,000 - f. 150,000	0.3 %
Category 12	f. 150,000 - f. 200,000	0.3 %
Category 13	f. 200,000 - f. 300,000	0.7 %
Category 14	f. 300,000 or more	0.7%
Category 15	don't know	31 %

## 4 Estimating the Bounds

In this section we apply the theory of Section 2 to the data on savings discussed in Section 3. We first estimate expressions (2) and (6) to show how bracket response can significantly improve Manski's (1995) original worst case set of bounds. We then examine the data to motivate the Weak Monotonicity condition in Section 2.2. Finally, we estimate the bounds under the three monotonicity assumptions.

The bounds in (2), (6), (10), (14) and (18) are functions of conditional expectations of observed quantiles and can be estimated using the available sample and, for example, nonparametric regression by means of kernel estimators (see for example Härdle and Linton, 1994). In our case, however, due to the small number of observations, we do not condition on any variable  $X$ , but instead use sample fractions to estimate the probabilities: since studies of the distribution of savings, income, etc., are usually expressed in terms of the quantiles, we use the estimated bounds to retrieve and report the bounds on the quantiles (see Section 2.3).

The distance between any upper and lower bound at each of the quantiles reflects

uncertainty due to item nonresponse; in order to measure uncertainty due to sampling error we place confidence bands around the estimated upper and lower bounds. Expressions (10), (14) and (18) involve estimation of  $\max[F_1, F(l)]$ . Analytic derivation of the asymptotic distribution of this estimator would be complicated; instead, we use a bootstrap method to find the confidence bands. This method consists on randomly re-sampling 500 times from the original data with replacement to estimate two-sided 95% confidence bands for both the upper and lower bound. We use the same bootstrap procedure to derive confidence bands for the estimates of the bounds in (2) and (6), although in these cases it would be straightforward to derive the pointwise asymptotic distribution. In the figures below, we report the upper confidence band for the upper bound: each point of the upper confidence band provides a 97.5% one-sided confidence band for the upper bound. Likewise, we report the lower confidence band for the lower bound. The (vertical) region between these two at each quantile shows an estimated interval that takes account of uncertainty due to both, sampling error and item nonresponse: with probability of at least 95% this region will contain the population quantiles of interest.

#### 4.1 Estimating Worst Case bounds

Figure 1 shows the estimated upper and lower bound for Manski's (1995) basic worst case bound where bracket information is not taken into account (expression (2)). The solid curves are the estimated upper and lower bounds whereas the dashed curves are the estimated upper and lower pointwise 97.5% confidence bands for each of the estimated bounds. We can see that the horizontal distance between estimated upper and lower bounds equals approximately 0.4, reflecting the initial percentage of item nonresponse.

**File Contains Data for  
PostScript Printers Only**

Table 3 gives a range of estimated quantiles with the corresponding pointwise confidence intervals corresponding to Figure 1. The distance between upper and lower bounds is very wide for any quantile. For example, with at least 95% confidence, the median is between  $f.350$  and  $f.42,000$ ; this width seems too large to be of practical relevance.

**Table 3: Estimated bounds and confidence intervals on savings (in Dutch Guilders) based on expression (2); Worst Case without bracket information.**

Quantiles	Confidence interval (Lower)	Lower bound	Upper bound	Confidence interval(Upper)
20 <sup>th</sup> Percentile	$f\ 0$	$f\ 0$	$f\ 2,400$	$f\ 3,500$
25 <sup>th</sup> Percentile	$f\ 0$	$f\ 0$	$f\ 3,806$	$f\ 5,400$
30 <sup>th</sup> Percentile	$f\ 0$	$f\ 0$	$f\ 5,935$	$f\ 9,122$
40 <sup>th</sup> Percentile	$f\ 0$	$f\ 0$	$f\ 11,929$	$f\ 19,958$
50 <sup>th</sup> Percentile	$f\ 350$	$f\ 800$	$f\ 26,725$	$f\ 42,000$
60 <sup>th</sup> Percentile	$f\ 1,500$	$f\ 2,400$	$f\ 200,000$	max
70 <sup>th</sup> Percentile	$f\ 4,345$	$f\ 6,000$	max	max
75 <sup>th</sup> Percentile	$f\ 6,413$	$f\ 9,300$	max	max
80 <sup>th</sup> Percentile	$f\ 9,300$	$f\ 11,850$	max	max
90 <sup>th</sup> Percentile	$f\ 20,000$	$f\ 27,093$	max	max

**File Contains Data for  
PostScript Printers Only**

Figure 2 shows the estimates of the worst case bounds where we include the information provided by the population of bracket respondents (expression (6)).

**Table 4: Estimated bounds and confidence intervals on savings (in Dutch Guilders) based on expression (6): Worst Case, Bracket information included.**

Quantiles	Confidence interval (Lower)	Lower bound	Upper bound	Confidence interval(Upper)
20 <sup>th</sup> Percentile	f 0	f 0	f 2,000	f 2,000
25 <sup>th</sup> Percentile	f 0	f 197	f 2,000	f 3,000
30 <sup>th</sup> Percentile	f 200	f 606	f 2,710	f 4,893
40 <sup>th</sup> Percentile	f 1,260	f 2,000	f 5,000	f 5,850
50 <sup>th</sup> Percentile	f 2,000	f 3,023	f 9,122	f 10,000
60 <sup>th</sup> Percentile	f 4,755	f 5,000	f 12,255	f 15,000
70 <sup>th</sup> Percentile	f 7,925	f 10,000	f 20,000	f 25,000
75 <sup>th</sup> Percentile	f 10,000	f 12,320	f 27,274	f 34,938
80 <sup>th</sup> Percentile	f 13,509	f 16,265	f 37,000	f 72,021
90 <sup>th</sup> Percentile	f 25,000	f 29,990	max	max

The interpretation of the curves in Figure 2 is similar to that of Figure 1. Comparing these two figures clearly shows how using the categorical questions can dramatically improve the knowledge provided by the bounds. Comparing Table 3 and Table 4 also shows the improvement of accounting for bracket information. For example, the width of the confidence interval for the median is reduced from f.41,650 in Table 3 to f.8,000 in Table 4; although the width is still large, the improvement is substantial.

## 4.2 Estimating Bounds with different assumptions of Monotonicity

Section 2 derives three bounding intervals according to three different assumptions of monotonicity. All three of them include the assumption  $F_1 \leq F_{01}$ ; we first motivate this assumption using the data of our empirical example. Although  $F_{01}$  is unknown we know that  $F_{01} \in [F(l), F(u)]$ ; we use the sub-samples of full respondents and bracket respondents to estimate  $F_1$ ,  $F(l)$  and  $F(u)$ . Figure 3 shows a plot of these estimates.

The solid step functions in Figure 3 are the estimates of  $F(l)$  and  $F(u)$  using the sub-sample of bracket respondents. The dashed curve is the estimate of the distribution function  $F_l$  using the sub-sample of full respondents. We see that, except at the very high quantiles of the

distribution,  $\hat{F}_1$  is in or below the area enclosed by  $[\hat{F}(l), \hat{F}(u)]$  suggesting that  $F_1 \leq F_{01}$ , except perhaps for very high  $y$ .

## File Contains Data for PostScript Printers Only

Table 5 gives the cumulative probability for various saving values; the column ‘full response’ corresponds to point estimates for  $F_1$  and the column ‘bracket response’ are point estimates of  $F(l)$ . Column 3 shows the results of testing whether the difference between each pair of probabilities is significantly different from zero;<sup>7</sup> under the null, this test statistic should be asymptotically standard normal. The results confirm that the distribution for full respondents is below that of bracket respondents except for very high values of  $y$ , and that the difference is significant for values from £5,000 to £40,000. Thus, Figure 3 together with the evidence in Table 5 motivate the weak monotonicity assumption.

Figure 4 shows the estimates for the bounds in (10), based on the weak monotonicity assumption. As in previous figures, the solid curves are the estimated upper and lower bounds on the quantiles of the distribution and the outside dashed curves are the estimated upper and lower confidence band for the upper and lower bound, respectively.

---

<sup>7</sup>The test is based on  $[\hat{F}(l) - \hat{F}_1]/\hat{\sigma}$  where  $\hat{\sigma}$  stands for the estimated standard deviation of  $[\hat{F}(l) - \hat{F}_1]$  such that  $\hat{\sigma}^2 = [\text{var}(\hat{F}(l)) + \text{var}(\hat{F}_1)]$ .

*Table 5: Frequencies for the first savings account and test for the difference between full respondents and bracket respondents.*

	Full response	Bracket response	Significance test
Units	455	207	
$\leq f$ 2,000 guilders	0.308 (0.022)	0.324 (0.033)	0.534
$\leq f$ 5,000 guilders	0.470 (0.023)	0.551 (0.035)	2.509
$\leq f$ 10,000 guilders	0.629 (0.023)	0.710 (0.032)	2.627
$\leq f$ 15,000 guilders	0.719 (0.021)	0.807 (0.027)	3.108
$\leq f$ 20,000 guilders	0.770 (0.020)	0.855 (0.024)	3.234
$\leq f$ 30,000 guilders	0.884 (0.015)	0.918 (0.019)	1.693
$\leq f$ 40,000 guilders	0.916 (0.013)	0.947 (0.016)	1.800
$\leq f$ 50,000 guilders	0.941 (0.011)	0.952 (0.015)	0.736
$\leq f$ 100,000 guilders	0.978 (0.007)	0.971 (0.012)	-0.713
$\leq f$ 150,000 guilders	0.987 (0.005)	0.976 (0.011)	-1.376
$\leq f$ 200,000 guilders	0.993 (0.004)	0.981 (0.010)	-1.889

**File Contains Data for  
PostScript Printers Only**

**Table 6: Estimated bounds and confidence intervals on savings (in Dutch Guilders) based on expression (10).**

Quantiles	Confidence interval (Lower)	Lower bound	Upper bound	Confidence interval(Upper)
20 <sup>th</sup> Percentile	f 0	f 0	f 1,348	f 1,750
25 <sup>th</sup> Percentile	f 0	f 197	f 1,878	f 2,400
30 <sup>th</sup> Percentile	f 200	f 606	f 2,400	f 3,755
40 <sup>th</sup> Percentile	f 1,260	f 2,000	f 4,990	f 5,935
50 <sup>th</sup> Percentile	f 2,000	f 3,023	f 8,955	f 10,000
60 <sup>th</sup> Percentile	f 4,755	f 5,000	f 12,255	f 15,000
70 <sup>th</sup> Percentile	f 7,925	f 10,000	f 20,000	f 25,000
75 <sup>th</sup> Percentile	f 10,000	f 12,320	f 27,274	f 31,530
80 <sup>th</sup> Percentile	f 13,509	f 16,265	f 37,000	f 66,136
90 <sup>th</sup> Percentile	f 25,000	f 29,990	max	max

Table 6 gives some of the quantiles in Figure 4. In this case the median of the distribution is between f.2,000 and f.10,000 with (at least) 95% confidence. Comparing this and other quantiles in Table 6 to those in Table 4 shows that imposing the weak monotonicity assumption does not lead to a great improvement compared to the worst case bounds; this is also clear from comparing figure 2 to 4.

The Monotonicity 1 and Monotonicity 2 assumptions involve  $F_{00}$  which cannot be retrieved from the data; we need additional information on the population of full non-respondents. One possibility is to look at various variables that could be related to wealth. The Center Data Panel provides information on ownership of cars, boats and other vehicles and on financial debts.

Table 7 subdivides the 755 individuals into the three sub-samples under study. The columns show the percentages of individuals that own the reported vehicles. The last column shows the percentage of individuals who have some form of financial debt with banks, a private financial institution, individual or retail companies. Furthermore, for the groups of full and partial savings respondents we break down the ownership rates by savings quantiles. The numbers in brackets are the standard errors for the estimated percentages. The table shows that for full non-respondents, the estimated ownership percentages are slightly higher than for the others, for all vehicles. This sub-sample also has the lowest percentage of financial debt. The rates per quantile suggest that ownership rates increase with savings, while debt holding falls with savings. Taken together, these findings suggest that individuals full non-respondents hold the highest amounts

**Table 7: Percentage (standard error) for various items of wealth for the sub-populations full respondents, bracket respondents and full non-respondents.**

	Units	Owners of cars	Owners of motorbikes	Owners of boats	Owners of caravans	Individuals with debts
$\delta_1=1$	455	<b>0.648 (0.022)</b>	<b>0.033 (0.008)</b>	<b>0.016 (0.006)</b>	<b>0.101 (0.014)</b>	<b>0.20 (0.019)</b>
Low 25%		0.536 (0.048)	0.018 (0.012)	0.0264 (0.015)	0.0720 (0.024)	0.324 (0.044)
25%-50%		0.632 (0.045)	0.0264 (0.015)	0.0088 (0.0087)	0.0720 (0.024)	0.212 (0.038)
50%-75%		0.684 (0.044)	0.0520 (0.021)	0.0264 (0.015)	0.088 (0.027)	0.148 (0.033)
High 25%		0.740 (0.041)	0.0352 (0.017)	0.0088 (0.0087)	0.176 (0.036)	0.116 (0.030)
$\delta_1=0, \delta_2=1$	207	<b>0.71 (0.032)</b>	<b>0.0435 (0.014)</b>	<b>0.058 (0.016)</b>	<b>0.092 (0.020)</b>	<b>0.26 (0.031)</b>
Low 25%		0.640 (0.066)	0.0384 (0.027)	0.232 (0.058)	0.368 (0.067)	0.272 (0.062)
25%-50%		0.680 (0.065)	0.0388 (0.027)	0.020 (0.019)	0.0760 (0.037)	0.212 (0.057)
50%-75%		0.792 (0.056)	0.0200 (0.019)	0.080 (0.038)	0.116 (0.044)	0.272 (0.057)
High 25%		0.772 (0.058)	0.0760 (0.037)	0.0388 (0.027)	0.136 (0.048)	0.288 (0.063)
$\delta_1=0, \delta_2=0$	93	<b>0.731 (0.046)</b>	<b>0.129 (0.035)</b>	<b>0.065 (0.025)</b>	<b>0.129 (0.035)</b>	<b>0.17 (0.039)</b>

of savings. This evidence would support the argument that when individuals are faced with a question on their savings, those who do not initially give an exact amount consist of two groups. On the one hand we have low savers who are not fully aware of the amount of their savings account; once they are given the chance to answer a range card question they will do so. The rest, who still do not disclose information about their savings, even in a categorical question, are those who will typically have high savings. They may refuse to reveal information about their savings, for example, because of confidentiality reasons. Thus, the above argument supports the Monotonicity 2 assumption leading to the bounds in (18). Figure 5 presents the estimates for these bounds.



## File Contains Data for PostScript Printers Only

Comparing Figure 5 with Figure 2 shows that bounds under Monotonicity 2 are sharper than the worst case bounds, particularly at the lower quantiles of the distribution. Table 8 compares point estimates of the worst case bounds with bracket information - column 3 - to bounds under the Monotonicity 2 - column 4 -. The third row in each cell shows the width between upper and lower bound. This comparison shows that monotonicity leads to an improvement for quantiles up to the 80<sup>th</sup> percentile of the distribution.

Finally, Figure 6 shows the consequence of estimating the bounding intervals based on Monotonicity 1 (expression (14)). Although Table 7 suggests that Monotonicity 2 is more plausible than Monotonicity 1, it is interesting to compare the estimates of the bounds under the two assumptions. Both Monotonicity 1 and Monotonicity 2 lead to narrower bounds than the Weak Monotonicity (compare Figures 4 and 5 and Figures 4 and 6). Monotonicity 2 leads to narrower bounds for the lower quantiles, while Monotonicity 1 improves precision at the higher quantiles.

*Table 8: Comparing point estimates from bounds estimated in (6) and bounds estimated in (18). Point estimates are based on 95% confidence bands.*

QUANTILES	POINT ESTIMATES	Point estimates of expression (6): Worts case and bracket response. (with 95% confidence)	Point estimates of expression (18). Bounds with Monotonicity 2. (with 95% confidence)
20 <sup>th</sup> Percentile	Lower bound:	fl. 0.00	fl. 325
	Upper bound:	fl. 2,000	fl. 1,750
	Difference:	<b>fl. 2,000</b>	<b>fl. 1,425</b>
25 <sup>th</sup> Percentile	Lower bound:	fl. 0.00	fl. 606
	Upper bound:	fl. 3,000	fl. 2,400
	Difference:	<b>fl. 3,000</b>	<b>fl. 1,794</b>
30 <sup>th</sup> Percentile	Lower bound:	fl. 200	fl. 1,196
	Upper bound:	fl. 4,893	fl. 3,755
	Difference:	<b>fl. 4,693</b>	<b>fl. 2,559</b>
40 <sup>th</sup> Percentile	Lower bound:	fl. 1,260	fl. 2,000
	Upper bound:	fl. 5,850	fl. 5,935
	Difference:	<b>fl. 4,590</b>	<b>fl. 3,935</b>
50 <sup>th</sup> Percentile	Lower bound:	fl. 2,000	fl. 3,676
	Upper bound:	fl. 10,000	fl. 10,000
	Difference:	<b>fl. 8,000</b>	<b>fl. 6,324</b>
75 <sup>th</sup> Percentile	Lower bound:	fl. 10,000	fl. 11,232
	Upper bound:	fl. 34,938	fl. 31,530
	Difference:	<b>fl. 24,938</b>	<b>fl. 20,298</b>
80 <sup>th</sup> Percentile	Lower bound:	fl. 13,509	fl. 14,825
	Upper bound:	fl. 72,021	fl. 66,136
	Difference:	<b>fl. 58,512</b>	<b>fl. 51,311</b>
90 <sup>th</sup> Percentile	Lower bound:	fl. 25,000	fl. 27,000
	Upper bound:	max	max
	Difference:	<b>not defined</b>	<b>not defined</b>

File Contains Data for  
PostScript Printers Only

## 5 Conclusions

In this paper we have extended the approach of Manski (1994, 1995) to deal with item nonresponse in micro surveys. Manski proposes to estimate bounds around the unknown conditional distribution function of the variable of interest. This approach does not fully identify the unknown distribution function. It avoids making additional assumptions on the data generating process. The extension in this paper consists of deriving bounds taking into account that initial non-respondents can sometimes provide partial information on the variable of interest. This is the case when they are routed to questions of a categorical nature, such as range card or unfolding brackets questions. Using the bracket information from these categorical questions can improve the bounds since they allow initial non respondents to provide information in the form of brackets. We derive and compute bounding intervals of a worst case type for the quantiles of savings in a Dutch cross section. For this variable the initial nonresponse rate approximates 40%. Once non- respondents are faced with the choice to provide information in the form of direct bracket response, the percentage of full nonresponse is reduced to 12.3%. Accordingly, we find much narrower worst case bounds if we take the brackets information into account. We also derive bounds that make use of several monotonicity assumptions; because we are dealing with three sub-populations - full respondents, bracket respondents and full non-respondents - there are many different monotonicity assumptions that can be made. We consider three of them, and interpret them using two different reasons for nonresponse: lack of information and concerns about confidentiality. We investigate the information available in the data to select, derive and estimate two bounding intervals under the concept of monotonicity which are in line with our data.

## Reference

- Härdle, W., and O. Linton, (1994), *Applied nonparametric methods*, in: *Handbook of Econometrics IV*, R.F. Engle and D.L. McFadden (eds.), North-Holland, Amsterdam, 2297-2339.
- Heckman, J.J., (1979), Sample selection bias as a specification error, *Econometrica*, 47, 153-161.
- Heckman, J.J., (1990), Varieties of selection bias, *American Economic Review, Papers & Proceedings*, 80, 313-318.
- Hurd, M., D. McFadden, H. Chard, L. Gan, A. Merrill, and M. Roberts, (1997), Consumption and savings balances of the elderly: Experimental evidence on survey response bias, *N.B.E.R., Conference of economics of Aging, working paper*.

- Juster, T., and J.P. Smith (1997), Improving the quality of Economic data: Lessons from the HRS and AHEAD, *Journal of the American Statistical Association*, 92, 1268-1278.
- Manski, C.F., (1989), Anatomy of the selection problem, *Journal of Human resources*, 24, 343-360.
- Manski, C.F., (1990), Nonparametric bounds on treatment effects, *American Economic Review, Papers & Proceedings*, 80, 319-323.
- Manski, C.F., (1994), *The selection problem in: Advances in Econometrics*, C. Sims (ed.), Cambridge University Press, 143-170.
- Manski, C.F., (1995), *Identification problems in the social science*, Harvard University Press.
- Nyhus, E.K., (1996), The VSB-CENTER Savings Project: Data collection methods, questionnaires and sampling procedures, *VSB progress report*, 42, Center for Economic Research (CentER), Tilburg University.
- Vazquez, R., B. Melenberg and A. van Soest, (1999), Nonparametric bounds on the income distribution in the presence of sample nonresponse, CentER DP paper 9933, Center for Economic Research (CentER), Tilburg University.

## Appendix A: Monotonicity Assumptions

### Block A1

$F_1 \leq F_{00} \leq F_{01}$	$F_1 \leq F_{01} \leq F_{00}$	$F_{01} \leq F_1 \leq F_{00}$
$F_1 \geq F_{00} \geq F_{01}$	$F_1 \geq F_{01} \geq F_{00}$	$F_{01} \geq F_1 \geq F_{00}$

Complete ordering of the distribution function for the three sub-populations

### Block A2

$F_{00} \geq F_{01} \text{ and } F_{00} \geq F_1$
$F_{01} \geq F_{00} \text{ and } F_{01} \geq F_1$
$F_1 \geq F_{00} \text{ and } F_1 \geq F_{01}$
$F_{00} \leq F_{01} \text{ and } F_{00} \leq F_1$
$F_{01} \leq F_{00} \text{ and } F_{01} \leq F_1$
$F_1 \leq F_{00} \text{ and } F_1 \leq F_{01}$

This table displays all cases where the maximum or the minimum of the three functions is known, but where no assumption is made on the relationship between the remaining two.

### Block B

$F_1 = F_{01} = F_{00}$		
$F_1 = F_{01} \leq F_{00}$	$F_1 = F_{00} \leq F_{01}$	$F_{01} = F_{00} \leq F_1$
$F_1 = F_{01} \geq F_{00}$	$F_1 = F_{00} \geq F_{01}$	$F_{01} = F_{00} \geq F_1$

This block displays all relationships involving an assumption of equality. The first corresponds to the exogenous selection assumption.

### Block C

$F_1 = F_{01}$	$F_1 = F_{00}$	$F_{01} = F_{00}$
$F_1 \leq F_{01}$	$F_1 \leq F_{00}$	$F_{01} \leq F_{00}$
$F_1 \geq F_{01}$	$F_1 \geq F_{00}$	$F_{01} \geq F_{00}$

This block displays all single relationships between two of the three functions. Nothing is assumed on the missing third.

